Big Data: Using ArcGIS with Apache Hadoop

David Kaiser  @ddkaiser
Michael Park

Esri DevSummit 2013

Session Offering ID: 301
Follow along with this presentation

http://esri.github.com/gis-tools-for-hadoop/devsummit2013

Use WiFi network: Esri2013

Big Data: Using ArcGIS with Apache Hadoop
GIS Tools for Hadoop

Hadoop users often have data with spatial value, but with limited options for spatially analyzing this data.

Esri has released an open-source framework to enable spatial-data processing in your Hadoop applications.

This enables you, as a developer, to build analytical tools that use both Hadoop and ArcGIS.
Why is this important?

Your Hadoop applications can provide spatial analysis

...and your users can leverage your Hadoop applications from within the ArcGIS Geoprocessing environment

Who is a Data Scientist?
Who is a Data Scientist?

Finding Your Data Scientist
Finding Your Data Scientist

Data scientists?

- engineering
- math and statistics
- comp sci
- hacking

awesome nerds

www.hilarymason.com
GIS Tools for Hadoop

The Hadoop 'Tools' are a combination of custom Hadoop applications and ArcGIS GP Tools.

- Sample tools that demonstrate the GIS capabilities provided for Hadoop
- Templates that can be used to build tools that solve specific big data problems

GIS Tools for Hadoop

- Esri Geometry API for Java
  - Java
  - Low-level geometric functions
  - Not necessarily specific to Hadoop

- Spatial Framework for Hadoop
  - Java
  - Utilities for reading JSON
  - Hive UDFs

- Geoprocessing Tools for Hadoop
  - Python
  - GP tools that copy to/from HDFS
  - GP tools that submit workflow jobs
  - GP tools that convert to/from JSON

Hadoop System

ArcGIS Desktop/Server

Developing Custom MapReduce Applications
# Geometry API for Java

**Simple API Functions for Java**

`com.esri.core.geometry.*`

<table>
<thead>
<tr>
<th>Relationship Analysis</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>equals</td>
<td>buffer</td>
</tr>
<tr>
<td>disjoint</td>
<td>cut</td>
</tr>
<tr>
<td>touches</td>
<td>clip</td>
</tr>
<tr>
<td>crosses</td>
<td>convexHull</td>
</tr>
<tr>
<td>within</td>
<td>intersect</td>
</tr>
<tr>
<td>contains</td>
<td>union</td>
</tr>
<tr>
<td>overlaps</td>
<td>difference</td>
</tr>
</tbody>
</table>
Spatial Framework for Hadoop

Enables developers to:
- spatially enable MapReduce applications

Enables Hadoop users to:
- run spatial Hive queries with ST_Geometry functions

Provides Java API's for:
- JSON Utility classes
- Hive UDF's

Uses the Geometry API for Java
Spatial Data in Hadoop

JSON files store collections of 'features'
- **Unenclosed JSON** is the dominant style; simple and appendable
- **Enclosed JSON** can optionally be used as a 'feature class'
  (A collection that should be analyzed as a complete set)

Accessing geometries from Hadoop Data Sources
- **com.esri.hadoop.json** - access JSON data as arrays of 'features'
- **com.esri.core.geometry** - construct geometry from arguments

Developing Custom MapReduce Apps

**Simple MapReduce**

```java
void map() {
  Feature feature = featureResult.getFeature();
  if (isFeature(feature)) {
    // Extract geometry
    Geometry point = new Point(...);
    // Process the geometry
    // ...}
```

**Hadoop**

- "features" is the dominant style; simple and appendable
- Optionally be used as a 'feature class'
- Could be analyzed as a complete set

**Hadoop Data Sources**
- access JSON data as arrays of 'features'
- construct geometry from arguments

**Eclipse Hadoop Wiki**

- JSON-Tools for Hadoop
- https://github.com/Ever/gis-tools-for-hadoop

**Demo**

JOIN earthquakes
WHERE ST_Contains (co, ST_Point (earthquakes.longitude, earthquakes.latitude))
GROUP BY counties.name
ORDER BY cnt desc;
https://github.com/Ever/gis-tools-for-hadoop
Simple MapReduce Code Sample

```java
void setup()
{
    inputStream = hdfs.open(new Path(config.get("input")));
    featureClass = EsriFeatureClass.fromJson(inputStream);
}

void Map(Long key, Text value)
{
    float longitude = Float.parseFloat(values[COL_LONG]);
    float latitude = Float.parseFloat(values[COL_LAT]);
    Geometry point = new Point(longitude, latitude);

    for (EsriFeature feature : featureClass.features)
    {
        if (GeometryEngine.contains(feature.geometry, point))
        {
            String name = feature.attributes.get(LABEL_ATTR);
            context.write(new Text(name), data);
            found = true;
            break;
        }
    }
}
```
ST_Geometry in Hive

SELECT counties.name, count(*) cnt FROM counties

JOIN earthquakes

WHERE ST_Contains (counties.boundaryshape,
ST_Point (earthquakes.longitude, earthquakes.latitude))

GROUP BY counties.name
ORDER BY cnt desc;

Geoprocessing Tools for Hadoop

Features To JSON, JSON To Features
- Provide serialization to and from JSON formats

Copy To HDFS, Copy From HDFS
- Moves files between ArcGIS and Hadoop

Execute Workflow
- Starts a workflow using the Hadoop Oozie workflow engine

Demo
Download the GIS Tools Project

Clone or Fork the project from Github
http://github.com/Esri/gis-tools-for-hadoop

Pre-built samples in the 'samples' directory

Place your completed tools in the 'tools' directory if you want to share them
Get the Source Code

Geometry API
http://github.com/Esri/geometry-api-java

Spatial Framework for Hadoop
http://github.com/Esri/spatial-framework-for-hadoop

Geoprocessing Tools for Hadoop
http://github.com/Esri/geoprocessing-tools-for-hadoop

To build the source:
ant
Contributing Your Work

Fork the gis-tools-for-hadoop project
- Hack on the code
  - Make new tools
    - Do awesome spatial analysis on big data
- Send a Github 'pull request' so we can pull the tool back into our project

Let us know how you are doing
- Add your content to the Wikis on Github
- Troubles? Open new Issues in Github
We want your feedback!

Session Feedback
http://esriurl.com/survey
Session Offering ID: 301

E-mail:
David Kaiser <dkaiser@esri.com> @ddkaiser
Michael Park <mpark@esri.com>
GIS Tools for Hadoop

The Hadoop 'Tools' are a combination of custom Hadoop applications and ArcGIS GP Tools.

Getting Started with Hadoop

GIS Tools for Hadoop

Contribute Your Work

Getting the Source Code

Download the GIS Tools Project

Finding Your Data Scientist

Developing Custom MapReduce Apps

Simple MapReduce Code Sample

ST_GeometryInHive